Semantic Search Engine for Data Management and Sustainable Development: Marine Planning Service Platform

Giuseppe Manzella, Roberto Bartolini, Franco Bustaffa, Paolo d'Angelo, Maurizio de Mattei, Francesca Frontini, Maurizio Maltese, Daniele Medone, Monica Monachini, Antonio Novellino, et al.

Semantic Search Engine for Data Management and Sustainable Development: Marine Planning Service Platform

Manzella, G., Bartolini, R., Bustaffa, F., D'Angelo, P., De Mattei, M., Frontini, F., Maltese, M., Medone, D., Monachini, M., Novellino, A., Spada, A.

INTRODUCTION

The value of marine environmental data is very high for the cost inherent to data collection, as well as for the knowledge that can be extracted from them. It is estimated that the EU Member States spend approximately € 1.85 billion a year on marine research (European Commission, 2010). About half is on infrastructure for facilitating observation. This includes ships, underwater observatories, floating buoys, drifting devices, remotely operated or autonomous underwater vehicles, as well as many other platforms, all equipped with a range of sensors and analytical capabilities. Unfortunately, a significant amount of data has been lost and is being lost for many reasons.

The NOAA National Data Buoy Centre reported that about 15-20% of data are lost for vandalism to buoys. However, the major quantity of data have been lost for the changes in recording technologies, that have affected the data integrity and rescue. Before electronic computers came into general use, oceanographic data were recorded in manuscripts, data reports, and card index files. With the advent of electronic data storage, oceanographic observations were increasingly recorded on magnetic media such as tapes and disks. Unfortunately, all these media are subject to degradation over time with subsequent loss of unique data. This has occurred in some cases, but unfortunately technology turnover is not the only reason for data loss. There are still a lot of researchers that are not 'publishing' their data in data centre systems, and with the result that data is lost when the researcher retires. Unfortunately large amounts of research funds are spent every year, while already existing data remain underutilised.

It has been underlined that data from the marine environment is a valuable asset; use and re-use can address threats to the marine environment, and can be used for the development of policies and legislation to protect vulnerable areas of our coasts and oceans, in understanding trends and in forecasting future changes. More in general, better quality and more easily accessible marine data can support the 'blue growth' or, in other words, the further sustainable economic development.

Data is analysed, synthesised, interpreted and transformed into information, and, as a final step, can produce knowledge. The outcome of this process is published as a scientific article. The intangible value of the data has pushed public authorities and organizations to encourage free and open access to data. In 2003 the 'Berlin Declaration on Access to Knowledge in the Sciences and Humanities' was published in order to "promote the Internet as a functional instrument for a global scientific knowledge base and human reflection and to specify measures which research policy makers, research institutions, funding agencies, libraries, archives and museums need to consider." (UNESCO, 2013). To make data usable in a tangible way it is necessary to accompany data with documentation, i.e. protocols, reports, grey literature published papers.

There are many reasons limiting the open and free access to data and documents, among which the Intellectual Property Rights and Copyrights. The idea of universal access to research, education, and culture is made possible by the Internet, but existing legal and social systems don't always allow that idea to be realized. To achieve the vision of universal access the Creative Commons (Clarke, 2001) is trying to create a balance between the reality of the Internet and the reality of copyright laws (Clarke, 2005).

The Marine Planning and Service Platform (MAPS) project started from a schematic depiction of the flow from research to library resources that is interlinking documentation and their underlying data. Within the project 'documentation' intend protocols and reports, as well as grey literature and papers published in conventional scientific journals. MAPS has developed a web search engine where the information retrieval is obtained from metadata and full text indexing and the information allow to select the underlying data from a database. With the existing information systems it is easy to connect the data with publications, and provide information on technologies used for data acquisition, laboratory analysis tools, protocols (tools and

technologies being, in a wider sense, part of those cultural artefacts developed in a certain historical period by scientists.

BACKGROUND

The accessibility to information, management of knowledge and its dissemination to the public, has a long history. However, there were, and still there area, many weak points in the transmission of knowledge among scientific communities and from scientific communities to the public.

In 1945 Wannever pointed out: 'Professionally our methods of transmitting and reviewing the results of research are generations old and by now are totally inadequate for their purpose. If the aggregate time spent in writing scholarly works and in reading them could be evaluated, the ratio between these amounts of time might well be startling. ... Mendel's concept of the laws of genetics was lost to the world for a generation because his publication did not reach the few who were capable of grasping and extending it; and this sort of catastrophe is undoubtedly being repeated all about us, as truly significant attainments become lost in the mass of the inconsequential.' The main statement of Wannever was essentially containing the idea behind the MAPS project: 'A record if it is to be useful to science, must be continuously extended, it must be stored, and above all it must be consulted.'

Many authors have underlined that knowledge can be tacit or explicit (e.g. Polanyi, 1966, Brown & Duguid 1998). The tacit knowledge is sometimes referred as 'know-how', while the explicit knowledge is sometimes referred as know-what (Brown & Duguid 1998). The former refers to codified knowledge, such as that found in documents, while the latter refers mainly to personal/experience-based knowledge.

In the MAPS concepts, tacit knowledge is part of the cultural environment of a person and the belonging to a community of practice. Although it is difficult to transfer to another person the tacit knowledge by means of (e.g.) writing it down, it is possible to built a knowledge management system were different types of knowledge can be discussed in some way. The personal contacts, people interactions and social media are means to transfer knowledge. Manzella and Manzella (2015) have carried out an experiment in a classroom with students having different cultural background (biologist, engineers, physicist, geologist), each having initial different personal beliefs deriving from their specific studies. They have been guided towards publications and to work with data, giving their own interpretation. Discussions on the different ideas allowed clarification of issues, reduction of uncertainties and shared understanding. However, it is a belief of the corresponding author of this chapter, that the tacit knowledge cannot be entirely transmitted. Any other person will elaborate the experiences of others and, after a knowledge building process, acquire intimately concepts, but in different forms.

The technological developments linked to computers and internet have offered the necessary tools for recording data, publications and any cultural artefacts and make them available to any user. The interlink between documentation and data needs two main components: an information system managing data and metadata and a semantic search system that extract structured information from a text. MAPS started from the consideration that a further development of a knowledge building system can support the social dimension of the sustainable development, providing tools to 'raise the level of understanding and commitment to action on the part of individuals, ... organisations ...' (Bruntland Report, WCED, 1987).

Information retrieval and semantic search

Information retrieval is the activity of obtaining user needed information from a collection of information resources. The idea of using computers to search for relevant pieces of information has practically started from the end of the second world war. In 1992 there was the first Text Retrieval Conference that catalyzed

research on methods that scale to huge corpora. The introduction of web search engines has boosted the need for very large scale retrieval systems even further. A web search engine is a software system that is designed to search for information on the World Wide Web. The information may be a mix of web pages, images, and other types of files. Some search engines also mine data available in databases or open directories. Unlike web directories, which are maintained only by human editors, search engines also maintain real-time information by running an algorithm on a web crawler. The currently popular search systems, including those implemented by common search engines, are mainly based on matching of strings: the words of the query are treated as a set of key terms whose presence or absence is sought in previously indexed texts (e.g. Brin, 1998; Kogut et Holmes, 2001).

Semantic search systems are slowly spreading amongst the generalist search engines. They consider context of search, location, intent, variation of words, synonyms, generalized and specialized queries, concept matching and natural language queries to provide relevant search results. In general, a preliminary analysis of the texts in question by means of existing technologies relying on lexical and conceptual generic resources is always necessary for the realisation of a statistical semantic search system in a specific domain. Later on, the results of this analysis – possibly manually corrected - are used for extracting new knowledge through specific terminology that will enrich the lexical resources: in this way these resources will act as a training set for the linguistic analysis systems. In the environmental science, a semantic search system was developed in the EU – FP7 project KYOTO (Knowledge Yielding Ontologies for Transition-based Organization), had the goal of providing a system for semantic search in order to allow expert users to model and improve their domain ontology with terms and concepts automatically extracted (Vossen et al, 2008).

Text Mining

The term Text Mining (TM) are all those semantic technologies that aim to extract structured information from unstructured data (Ramjan et al., 1998).

The main applications of Text Mining and Information Retrieval technologies are the search systems that allow to retrieve texts or portions of text contained in a document base through a string of natural language search, carried out by the user with a simple query.

Research systems commonly in use, including those implemented by common search engines, are based primarily on matching strings: the words of the query are treated as a set of key terms whose presence or absence is sought in the texts previously indexed.

There are several statistical algorithms to determine the similarity of the query text and one contained in the documents (Lott, 2012). It is important to note, however, that such algorithms do not make a semantic analysis of the content of the query and even the content of the documents under consideration. Such systems therefore have many problems given the well-known ambiguity of natural language.

For this reason, the research in this field is pushed toward systems that are able to understand and structure, at least in part, semantic information contained in the text. In semantic search, words are not regarded as mere strings, but as a unit with meaning that can be reported by the context in which they are located; this allows to identify the concepts and semantic relationships that are expressed in the words of the query and go and see if these concepts or relationships (or concepts / relationships like) are present in the texts in question. Simultaneously, the system allows to ignore and discard only superficially different concepts that are represented by the same strings. To operate the system must be able to make the semantic disambiguation of queries is that the words contained in the texts. In this context it is essential to have access to knowledge bases that identify related concepts.

If more simple research systems semantics are limited to recognize and manage homonymy and synonymy, the most sophisticated one want to extract complex levels of information, such as (i.e.) 'events'. For the extraction of such levels of information is necessary to recover the semantic relationships between terms, from grammatical relations, the events, the participants in the action and their semantic roles.

Information system

During the last decades important marine data infrastructures have been developed at international level. Good examples are provided by the IOC-UNESCO IODE Ocean Data Portal (Belov, S., Mikhailov N., 2012), the Australian Integrated Marine Observing System (Proctor et al, 2010), the European SeaDataNet (Maillard et al, 2007), the US NOAA National Oceanographic Data Centre (Boyer et al., 2013) and many others.

These information systems have the primary goal to provide transparent access to marine data sets and data products. They have developed standardized distributed systems for managing large and diverse datasets (i.e. temperature, salinity current, sea level, and many other chemical, physical and biological properties) collected by different sensors on different platforms. ISO and OGC compliant discovery services and metadata models have been adopted in order to assure interoperability.

To make scientific web portals effective elements of knowledge management, the information systems should be linked to the semantic search systems (and vice versa) by providing the necessary information on geographical and temporal coverage, parameters. This is the goal of MAPS project.

SPECIFIC ONTOLOGY BASED SEMANTIC SEARCH ENGINE

Behind the MAPS project, there is the idea of implementing a 'knowledge management' system as a basic element for the wider objective of participation to societal advances in science, technology and sustainable development. The knowledge management system is based on two main components:

1. knowledge building environment - a continuous learning practice that allows the presentation of theories, the comparison with observations, understanding and resolution of controversies
2. service and planning platform - an open and free access to data/products and information related to the data acquisition (including environmental conditions), instrumentation, protocols for in situ and ex situ practices used for quality assessment and quality control.

The knowledge management model proposed is represented in Figure 1, where the knowledge building element is strongly linked to the information system, that is providing access to data and products, and must include also links to documents.

Figure 1. MAPS concept showing the interlink between documentation and data

A portal managing at the same time both data and documentation allows any user to find and assess scientifically credible information about the (e.g.) the marine environment.

In general, the realization of a semantic search system in a specific domain requires a first analysis of texts based on technologies that rely on conceptual generic lexical resources. This first analysis is used to extract new knowledge through specific terminology: in this way there will be an adjustment of the systems of linguistic analysis to lexical resources. This 'linguistic annotation' is added to the text information that aim to make explicit the implicit structured information in the document. The annotation can be done manually,

but generally for large amounts of texts exist linguistic analysis tools that automatically write down the various levels that are often dependent on each other, as some are preparatory to the identification of others.

The tools of linguistic analysis can be ‘language specific’, if only one language is used, or language independent, if more than one languages are used. Typically the tools that work on more than one language extract information of the highest level and lean annotation lower levels (for example morphosyntactical). Without going into detail, the automatic language analysis tools fall into two broad categories:

1. Tools that implement in their code explicit rules ("rules");
2. Statistical tools.

Statistical tools that implement algorithms in a supervised machine learning (with a corpus of training annotated by hand) or unsupervised (inductively) can "learn" to recognize the regularities in the text. For a state of the art tools on linguistic see among others Parra et al 2009. For the importance of record in semantic search see also Uren et al., 2006.

The different levels of linguistic annotations are:

- Language Detection: is the process of association of a generic text to the language in which it was written. Typically it is based on the calculation of the number of stop words (prepositions, article, words from classes closed) within the text.
- Tokenization is the process of dividing a text in minimum units of analysis called tokens, eg words, punctuation, dates, initials etc. The process of Tokenization can be very complex for some languages and relatively simpler for others, for example in Italian, thanks to the presence of spaces and stop word, the process is quite easy.
- Lemmatisation: is the process of reducing the inflected form of a word to its canonical form (unmarked), called lemma. The standardized form is the gateway to the language vocabulary. The presence of linguistic ambiguity means that the association is typically one-to-many.
- Analysis Morphology and / or Part-of-speech tagging: consists in associating with each word of the text to its Part-of-Speech and possibly other morphological features such as the kind, number, mode and time for verbs. If first phase lemmatisation the task is reduced to morphological disambiguation. Generally the approach to this task and statistical or methods are used supervised learning machine learning which allow accuracies around 95%.
- Shallow Parsing: is the subdivision of a text phrases for analyzing morphosyntactic, or in blocks nominal, prepositional, verbal etc. (If these elements are flat, without presenting a branched structure, non-recursive, they are called chunks) and represent a linguistic unit in large granularity: for example a nominal chunk consists of the name preceded by 'any article, prepositional chunks are introduced by preposition more head carrier of lexical meaning.
- Deep Parsing: with this kind of deep analysis means the task of identifying dependencies between the functional portions of text, such as dependencies "be subject" or "be" a verb, than the phase of surface analysis that implicitly identified only relationships between words.
- Named Entity Recognition and Classification (NERC): consists in 'assign a semantic category in a sequence of words close, perceived as a concrete entity par. The state of the NERC currently allows the identification and classification of high-level categories such as people, locations, organizations, and artifacts.
- Coreference recognition (COREF): binds to the Named Entity Recognition (NER) since it allows to recognize portions of text that coreference to the same entity, for example pronouns or other; in this way it is possible to recover property or event in which such entity is also involved because it is not named in full.

- Word Sense Disambiguation (WSD) consists in 'assignment of a proper contextual meaning (sense) to a particular word in the text or speech. Many international research groups are working to WSD, using a variety of approaches. However, to date, there is no record of accurate systems of WSD with broad coverage [28]. Currently the accuracy achieved for English texts with trained systems on a limited number of words is 60-70% (due to the large effort required for the manual annotation of examples). The WSD is still one of the most important open problems in NLP.
- Semantic Role Labeling (SRL) is the assignment of the appropriate relationship between a predicate and its syntactic constituents. Typical topics include semantic agent, patient, instrument, subject added, complementary rental, time, mode, etc. cause. Recognize and label semantic arguments is a fundamental task to answer the following questions: "Who", "When", "What", "Where", "Why", as they are fundamental issues to be solved in the areas of Information Extraction (IE), Question Answering (QA) and Text Summarization (TS).

Currently, there are a number of tools for the analysis of natural language, but given the fragmentation of the methodologies and technologies is impossible to refer to a single source for a complete list; as regards the English and other European languages :

- TreeTagger: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
- Freeling: http://nlp.lsi.upc.edu/freeling/
- openNLP: https://opennlp.apache.org/index.html
- Stanford CoreNLP: http://nlp.stanford.edu/software/corenlp.shtm
- Morph-it: http://dev.sslmit.unibo.it/linguistics/morph-it.php

To evaluate a tool, the scientific community provides ‘gold standard’ (benchmarks), manually annotated corpora to calculate precision / recall and accuracy. In addition to accuracy it is also important usability of the instrument, the possibility of access to the code to change the rules or to improve the training with new data. For this reason, the tools to be favored are those distributed as open source.

It is important to note that many of the tools available freely online cannot be used for commercial purposes, but only for research purposes.

In MAPS have been used the tool suite developed for the project (opener) Open Polarity Enhanced Name Entity Recognition. For a description of the project and tools made from Apache OpenNLP, whose license also allows commercial use, see the section on the relevant projects.

SPECIFIC DATA ACCESS INFRASTRUCTURE

Any observing or modeling system needs a robust data and information system, capable of combining data and knowledge gathered over time with new observations and modelling results to provide a range of integrated, interdisciplinary datasets, indicators, visualizations, scenarios, and other information products. The present efforts on data preservation and information systems include multiple sources of information, involve multiple stakeholders, support effective decisions at global to local scales, provide full and open access to data (Diviacco and Leadbetter, 2017, chapter 2 this book).

The information system is the real interface for most users, providing the data, information and products required to support research and address societal decisions in diverse areas such as climate studies and adaptation, disaster warning and mitigation, and ecosystem-based management. Beyond the observation elements there are many modelling, data assimilation and synthesis activities that provide added value to observations and meet specific user requirements for information.

Providing open access to data and related products reduces duplication among the user community and promotes data standards and broad accessibility This supports the principle of "measure once/use many times.

The Physics portal (www.emodnet-physics.eu) of the European Marine Observation and Data Network (EMODnet) has been used to provide both real time and historical data, using the experience done within the SeadataNet (www.seadatanet.org) network of National Oceanographic Data Centers. EMODnet Physics provides a single point of free and open access to marine real time and archived data on physical conditions of all European Seas as monitored by fixed platforms, ferry box, ARGO, gliders, etc.

The main elements derived from the SeaDataNet 'system' are the vocabularies and the Common Data Index (CDI), that can be considered a de-facto standard for marine metadata in Europe. The CDI format is a marine profile of the ISO19139 metadata content standard compliant with INSPIRE Directive Implementing Rules.

In MAPS, the information system has been a scientific field of study addressing both operational activities and the distribution and use of information to knowledge communities. This means that the Information system in MAPS case has been referred also to the interaction that can occur across organizational boundaries. In other words, the information system must include important elements of information management as well as knowledge management together. In practice, the system must be able to (e.g. Ward and Peppard, 2002):

- Manage and maintain data to be interpreted in order to render information
- The information has to be understood in order to emerge as knowledge
- Knowledge must be used by a wide variety of stakeholders, including decision makers
- Stakeholders use of knowledge is expected to deliver meaningful results.

It has been underlined by some authors that this approach (called DIKAR: Data, Information, Knowledge, Action and Result) is aligning technology and organisational strategies, and it can be seen as a strategic change in information management systems. The recognition that information management is an investment that must deliver meaningful results is important to all modern organisations that depend on information and good knowledge for their success.

SOLUTIONS AND RECOMMENDATIONS

In this paragraph are presented some important projects in which semantic search technologies and / or textual analysis have been developed for the extraction and retrieval of information from texts domain specific.

KYOTO

The project Yielding Kyoto Knowledge Ontologies for Transition-based Organization has aimed to provide a system for searching deep semantic content, and that would allow experienced users to model and improve its domain ontology with terms and concepts automatically extracted. As domain environment for Kyoto was chosen environmental sector. It has gone from a knowledge base consisting of predefined generic ontology (SUMO for example) connected to the WordNet of each language treated by the project; It is then used this lexical basis of conceptual knowledge both to allow the semantic analysis and research on a number of documents to extract new candidate information from such documents; eventually the system enables experienced users to evaluate and integrate the concepts and terms candidates in their knowledge base, improving the analysis of new documents.

- Link project site: http://kyoto-project.eu/xmlgroup.iit.cnr.it/kyoto/index.html

- Link tools: http://kyoto-project.eu/ xmlgroup.iit.cnr.it/ kyoto/ index6b4e.html
- Link deliverables: http://kyoto-project.eu/xmlgroup.iit.cnr.it/kyoto/index4160.html

GLOSS

The project GLOSS Multi Lingual Information Access to Multi Media Contents (MLIA2MMC) is an emerging technology and enabling that permit new forms of multilingual access to information, by combining the latest developments in Text Mining, Knowledge Engineering (IC), Language Processing Natural (TAL) and Semantic Interpretation (IS). The ability of these tools is to provide targeted and effective access to content: location of information in an easy to read grid enriched with functional claims (who does what), time (when), geographic (where). This resilience has become a key feature for the search engines of the future. The overall objective of the project GLOSS - derived from Kyoto - but addressed the geographical domain, is focused on graphical visualization tools. GLOSS The project was designed to carry out a research of industrial type starting with the current state of the art technology MLIA2MMC, thereby allowing the construction of an integrated environment for the consultation, the production and sharing of knowledge.

- Link with Project: http://weblab.iit.cnr.it/gloss/

PANACEA

The project PANACEA Platform for Automatic, Normalized Annotation and Cost- Effective Acquisition of Language Resources for Human Language Technologies (STREP – Specific Targeted Research Project - under EU-FP7) has developed a factory of linguistic resources (both data and algorithms) that automates all the steps involved in acquisition, production, maintenance and upgrading of resources required by the machine translation. The factory also includes other language technologies such as those inherent alignment of parallel corpora, the production of bilingual dictionary and the production of lexicons very informative. The reference domain is environmental / legal.

- Link project site: http://www.panacea-lr.eu/
- Link tools: http://www.panacea-lr.eu/en/info-for-professionals/the-platform/
- Link deliverables: http://www.panacea-lr.eu/en/project/work-packages/

Opener

The project Opener Open Polarity Enhanced Named Entity Recognition is funded by the European Commission under the 7th Framework Programme. The main goal of opener and provide a set of tools is immediately usable by researchers from small and medium enterprise to perform activities of natural language processing, free and easy to integrate into their workflow. Opener also wants to be able to detect and disambiguate entities mentioned and perform sentiment analysis and opinion on the texts detection, so you can capture information and steal the sentiment and market views about the products reviewed on the web.

- Link project site: www.opener-project.org
- Link tools: http://www.opener-project.eu/webservices
- Link deliverables: http://www.opener-project.eu/documentation/
- Link extraction multiwords: https://github.com/opener-project/multiword-generation

BootStrep

The project BooStrep Bootstrapping of Ontologies and Terminologies Strategic Research Project was funded within the 6th Framework Programme of the EC. The project - aimed at the biological realm - combined in a

common framework - terminology resources and existing databases, and has implemented a system of analysis of texts for the acquisition of new terms, concepts and relationships, or a semi-automatic process of construction ontologies in biology.

- Link project site: http://www.bootstrep.org/
- Link tools: http://www.bootstrep.org/resources.html
- Link to the result of the project: http://www.ebi.ac.uk/Rebholz-srv/BioLexicon/biolexicon.html

Information system

The MAPS information system was developed in a previous regional project (MARINE) and is constituted by a fully distributed system where data and metadata are collected and maintained by different project partners. The system is INSPIRE compliant and is using the following elements from the European project SeaDataNet:

- Common Data Index (CDI) indexing the data and thus being the element of the catalogue, allowing search, localization, selection, semantic interoperability, asset management and availability; CDI is containing also the links to documentation;
- BODC vocabularies;
- NetCDF and Ocean Data View formats

The services provided by the information system are Catalogue Service for data search, Web Map Service for consultation, Web Feature Service for downloading, Coordinate Transformation and Service Chain to claim other services (see for the general concepts Glaves, 2017, chapter 3 this book).

Each node of the federated information system is structured as follows:

- *PlatformList.xml* file is giving the list of platforms that have collected data made available though ftp protocol
- *PlatformsCDI.xml* file is giving information on each platform listed in *PlatformList.xml*
- A directory *latestData* is containing the data collected during the last 30 days; one file is containing daily data from one platform; daily subdirectory are continuously formed;
- A directory *monthlyData* is archiving the data older that 30 days; it is organised as the *latestData* directory;
- A directory *biblioData* is archiving documentation on data acquisition, protocols, reports; it is organised in subdirectories for the different platforms.

PlatformList.xml files have KML format and allow the easy management of geospatial data.

The MARINE information system used in MAPS is schematized in Figure 2.

Figure 2. The federated information systems used in MAPS project

The Common Data Index

The use of the Common Data Index (CDI) XML is of strategic importance in the project. It has been developed within the European projects SeaSearch, SeaDataNet and SeaDataNet2. However, some

additional elements have been implemented in MAPS in order make it more effective in combination with a search engine.

From the original CDI (www.seadatanet.org), MAPS has used those elements that allow the selection of important information, such as the parameter name (e.g.)

<gmd:keyword>
<sdn:SDN_ParameterDiscoveryCode codeSpace="SeaDataNet" codeListValue="TEMP"
codeList="http://.../schemas/SeaDataNet/Codelist/sdnCodelists.xml#SDN_ParameterDiscoveryCode">Temperature of the water column</sdn:SDN_ParameterDiscoveryCode>
</gmd:keyword>

or instruments (e.g.)

<gmd:keyword>
<sdn:SDN_DeviceCategoryCode codeSpace="SeaDataNet" codeListValue="13"
codeList="http://.../schemas/SeaDataNet/Codelist/sdnCodelists.xml#SDN_DeviceCategoryCode">bathythermographs</sdn:SDN_DeviceCategoryCode>
</gmd:keyword>

Some elements were better defined, such as the information on publications

<gmd:CI_Citation> <gmd:title>
<gco:CharacterString>Improved near real-time data management procedures for the Mediterranean ocean Forecasting System-Voluntary Observing Ship program</gco:CharacterString>
</gmd:title>
<gmd:date> <gmd:CI_Date> <gmd:date>
<gco:Date>2003</gco:Date>
</gmd:date>
<gmd:dateType>
<gmd:CI_DateTypeCode codeList="http://.../schemas/SeaDataNet/Codelist/gmxCodelists.xml#CI_DateTypeCode"
codeListValue="publication" codeSpace="ISOTC211/19115">publication</gmd:CI_DateTypeCode>
</gmd:dateType> </gmd:CI_Date> </gmd:date>
<gmd:citedResponsibleParty> <gmd:CI_ResponsibleParty> <gmd:individualName>
<gco:CharacterString>G. M. R. Manzella</gco:CharacterString>
</gmd:individualName>
<gmd:organisationName>
<gco:CharacterString>ENEA</gco:CharacterString>
</gmd:organisationName>
…
<gmd:role> <gmd:series> <gmd:CI_Series>
<gmd:name> <gco:CharacterString>Annales Geophysicae</gco:CharacterString> </gmd:name>
<gmd:issueIdentification> <gco:CharacterString>1</gco:CharacterString> </gmd:issueIdentification>
</gmd:CI_Series> </gmd:series>
<sdn:onlineResource> <gmd:CI_OnlineResource> <gmd:linkage>
<gmd:URL>http://www.ann-geophys.net/21/49/2003/angeo-21-49-2003.html</gmd:URL>
</gmd:linkage> </gmd:CI_OnlineResource> </sdn:onlineResource> </gmd:CI_Citation> </sdn:additionalDocumentation>
</gmd:MD_DataIdentification> </gmd:identificationInfo>

For the digital identifier (DOI) the NOAA (http://www.ngdc.noaa.gov/docucomp/page) elements have been used

<gmd:identifier> <gmd:MD_Identifier> <gmd:code>

<gmx:Anchor xlink:href="http://dx.doi.org/10.7289/V5154F01" xlink:title="DOI"
link:actuate="onRequest">doi:10.7289/V5154F01</gmx:Anchor>
</gmd:code> </gmd:MD_Identifier> </gmd:identifier>

As well as for right constraints

<gmd:resourceConstraints xlink:title="NGDC Data Citation Statement">
<gmd:MD_LegalConstraints> <gmd:useLimitation>
<gmd:MD_RestrictionCode codeList="http://.../schema/resources/Codelist/gmxCodelists.xml#MD_RestrictionCode"
codeListValue="otherRestrictions">otherRestrictions</gmd:MD_RestrictionCode>
</gmd:useLimitation>
<gmd:otherConstraints>
<gco:CharacterString>Cite as: Manzella, G. M. R., Scoccimarro, E., Pinardi, N., and Tonani, M.: Improved near real-time data management procedures for the Mediterranean ocean Forecasting System-Voluntary Observing Ship program, Ann. Geophys., 21, 49-62. doi:10.5194/angeo-21-49-2003, 2003</gco:CharacterString>
</gmd:otherConstraints> </gmd:MD_LegalConstraints> </gmd:resourceConstraints>

In this way, the Common Data Index has all the elements selected in MAPS for the semantic search.

Digital library

The library is containing the documentation and the search engine that is composed by:

- Syntactic parser - This module is responsible for the extraction of "rich words" from the text: the whole document gets parsed to extract the words which are more meaningful for the main argument of the document, and applies the extraction in the form of N-grams (mono-grams, bi-grams, tri-grams).
- MAPS database - This module is a simple database which contains all the N-grams used by MAPS (physical parameters from SeaDataNet vocabularies) to define our marine "ontology".
- Term extractor - This module performs the most important task of filtering the N-grams extracted from the text by the parser with the provided oceanographic terminology. It checks N-grams supplied by the syntactic parser and then matches them with the terms stored in the MAPS database. Found matches are returned back to the parser with flexed form appearing in the source text.
- A "relaxed" extractor - This option can be activated when the search engine is launched. It was introduced to give the user a chance to increase the ontology with new N-grams combining existing mono-grams and bi-grams in the database with rich-words found within the source text.

The innovation of a semantic engine lies in the fact that the process is not just about the retrieval of already known documents by means of a simple term query but rather the retrieval of a population of documents whose existence was unknown. The system answers by showing a screenshot of results ordered according to the following criteria:

- Relevance – of the document with respect to the concept that is searched
- Date - of publication of the paper
- Source – data provider as defined in the SeaDataNet Common Data Index
- Matrix - environmental matrices as defined in the oceanographic field
- Geographic area - area specified in the text
- Clustering – the process of organizing objects into groups whose members are similar

The clustering returns as the output the related documents. For each document the MAPS visualization provides:

- Title, author, source/provider of data, web address
- Tagging of key terms or concepts
- Summary of the document
- Visualization of the whole document

Library model

MAPS can be applied to documents in English and in Italian. The design of the semantic engine architecture is based on the integration of already existing linguistic/semantic modules emphasizing the adaptation to the lexical and terminological document bases of the specific marine domain. A general source of inspiration for MAPS's terminology base is WordNet (Fellbaum, 1998), a lexical database that groups English words into sets of synonyms, providing short definitions and recording a number of relations among these synonym sets or their members. WordNet can thus be seen as a combination of dictionary and thesaurus. The WordNet model was applied to Italian by Martinelli et al (2000), with the creation of ItalWordNet, and used in several projects to enhance semantic search in documents; Marinelli and Roventini (2006) enhanced ItalWordNet with maritime terms.

Search Engine

The MAPS system was conceived so as to receive objects of various nature (text and data) that users need to consult via a web interface. MAPS objects are also documents that need to be discovered in their content, as well as by means of their metadata. Figure 3 represents the search architecture adopted for MAPS. The documents are indexed and made searchable by the Semantic Engine and then queried from the web interface by users together with other types of objects (data).

In particular, the Search Engine uses semantic-conceptual technologies in order to extract key concepts from unstructured text such as technical documents (reports and grey literature) and scientific papers. Any text has to be made indexable and searchable by the end user in the same way as the structured data (such as oceanographic observations and metadata) are.

In order to achieve this purpose a semantic system has been implemented using natural language processing techniques and language engineering tools. They allow the correct indexing of documents and the correct processing of natural language queries in the human-machine search interface. This process is completed by incrementing the terminological knowledge base of domain concepts with new elements found in the texts.

Figure 3 – The semantic search system.

As shown in Figure 4, the Semantic system contains a Natural Language Processing (NLP) pipeline that analyses texts thanks to NLP tools and pre-existing terminological databases, both generic and domain specific. The result of the NLP pipeline is an annotated text that allows for terminological extraction of relevant concepts. Such terms are then used for indexing, but can also enrich a domain lexicon Data Base that can be later re-fed into the NLP pipeline.

The NLP and indexing processes are asynchronous and are activated whenever a new set of texts enters the document base.

Figure 4 - The semantic engine.

The query analysis pipeline is instead a synchronous process, that is called whenever the search engine is interrogated by a user with a query in natural language. For this reason the query analysis system is built in a simplified way that guarantees to obtain similar results to the more complex NLP and indexing pipeline (so as to ensure the extraction of relevant texts), but with better performances in terms of response time - at least

for such smaller snippets of texts that are likely to be entered by humans in a query. The logical steps for NLP and term extraction are shown in Figure 5.

Figure 5. Steps of Natural Language Processing and term extraction.

The first step is the transformation of the original document (often in pdf format) in plain utf-8 format text. This is done with an open source command-line utility (pdftotext) converting PDF files to plain text files - i.e. extracting text data from PDF-encapsulated files.

Subsequently, since documents may be in English or Italian, a language detector is used in order to call for the correct language specific NLP pipeline. This is done by means of the Google language detector cld2 (https://code.google.com/p/cld2/). Simplified ad-hoc versions for the two query pipelines have been implemented, using frequent terms in both languages. In practice, the logical steps for the NLP pipelines are the same for both languages, but some pre-existing tools are used whenever possible:

1. sentence splitting, dividing the text in sentences
2. word tokenization, splitting sentences into words
3. lemmatization and morphological analysis (part of speech tagging)
4. toponym detection, identifying geographic names
5. basic syntactic analysis (chunking) dividing the sentence into non recursive constituents

More specifically the component used for document in English are:

- For sentence splitting the SentenSentenceDetector of the Apache OpenNLP suite was used (https://opennlp.apache.org/)
- For tokenization, lemmatization and morphological analysis the Genia Tagger was adapted (http://www.nactem.ac.uk/tsujii/GENIA/tagger/)
- For toponym detection and identification of geographic names GeoNerD (http://sourceforge.net/u/geonerd/profile/) ad hoc develop method using a knowledge base (Geonames) and a rule based algorithm
- For syntactic analysis Chunker Eng, ad hoc developed module using a rule based algorithm (e.g. Wang, 2008)

The component used for document in Italian are:

- Freeling, an open source language analysis tool suite available for several languages among which Italian (http://nlp.lsi.upc.edu/freeling/).
- GeoNerD: ad hoc develop method using a knowledge base (Geonames) and a rule based algorithm.
- Chunker Ita, ad hoc developed module using a rule based algorithm (Lenci et al., 2003).

Terminology extraction

The NLP pipeline produces an intermediate annotated document, which is preliminary to terminology extraction; the latter in turn is necessary in order to be able to correctly index the document in the document base to be later semantically searched. The terminology extraction tool, Ideal, takes the chunked text as input, containing all required morphosyntactic information. Ideal is a rule based system and specific rules were developed for Italian and English. This tool was developed specifically for MAPS, being the most important part of the NLP pipeline. The rules, written in an ad hoc language (Bartolini et al, 2004), are designed to extract all simple and complex noun phrases in the text. The extractor works on the chunked text searching for patterns such as :

- nominal phrase;

- nominal phrase followed by one or more adjectival or prepositional phrases

All possible intermediary marches are produced, and they are taken to represent concepts of different degrees of specificity. Thus the algorithm not only returns candidate terms, but sub-trees of candidate terms linked by hierarchical relations. For instance a document containing the text mean sea salinity will produce the following "family" candidate terms:

- salinity (more generic)
- mean salinity (less generic)
- mean sea salinity (specific)

In particular each complex term is analysed into a lemmatized head term (salinity) and all the possible specifiers, with the frequency with which they occur in the text.

In order to guarantee that such terms are domain relevant terms a filtering operation is then performed. To this purpose a list of concepts is stored in the terminological base of the semantic engine, drawn from SeaDataNet vocabularies (see Buck and Lowry, 2017, chapter 1 this book). Whenever a set of related terms is extracted, it is compared to the list of terms in the terminological base. If one of the related terms or a sub component thereof is also present in the terminology base, the whole family of related terms is validated.

Filtering can be obtained by:

- exact matching: only the terms corresponding to those found in the terminological base are used for indexing;
- relaxed match: one of the related terms is also present in the TB, the whole family of related terms is validated.

So in the previous example, if salinity is present in the domain term base, in the case of exact matching only salinity is used for indexing the document, whereas in the case of relaxed matching salinity, mean salinity and mean sea salinity are used for indexing.

While the exact matching method may be used to increase precision in the semantic search, the second option is more interesting, as it allows for the discovery of new complex domain terms that are not already in the domain terminology. For instance from given the domain term data, it can be derived satellite data, wavemetrics data, oceanographic data; from wave it can be derived mean wave height, wave height variation, ...

The NLP and terminology extraction pipeline has a set of terms in a standardised json format as output, which is then used as input for the indexing function in the Apache Lucene Core (https://lucene.apache.org/core/)

Similarly, the query analysis pipeline produces the same output that can be used to search the existing index for matching terms.

The service platform

During the last stages of the MAPS project, a trial activity was carried out in order to assess the capabilities of the developed search engine. The trial activities were broken down in two phases:

- the first one was responsible of assessing whether stakeholders requirements were correctly implemented in the pilot system;

- the second phase had the objective to evaluate whether the search engine is capable of retrieving relevant documents.

The assessment of requirements is usually carried out by planning and executing a number of tests in order to verify the behaviour of the system in given operational conditions. This approach is usually adopted during user acceptance project phases, where stated requirements are verified. However, the trial was finalized more on evaluating the quality attributes of the implementation than verifying that every function was properly executing. As such, it was decided to evaluate the functionality of the pilot system by measuring the quality attributes defined by a software quality model versus stakeholders' requirements.

A software quality model is a collection of quality attributes that a piece of software should have. The quality of the developed software is then assessed by evaluating whether or not it possesses all the required attributes. Attributes are generally structured in two or more hierarchical levels so that it is easier to figure them out and verify the software against them.

There are several quality models available. The first software quality models were developed by McCall (1977) and Bohem (1978). More recently, ISO has developed a number of standards defining software quality models, the most known of which is the ISO/IEC 9126 standard. This model has been used for assessing the search engine.

The ISO 9126 quality model was based on the McCall and Boehm models. The model has two main parts consisting of:

- Internal and external quality attributes
- Quality in use attributes.

Internal quality attributes are referred to the system properties that can be evaluated without executing, while external quality attributes refer to the system properties that can be assessed by observing it during execution and are experienced by users when the system is in operation.

The quality in use attributes are referred to usage of the product and regards its effectiveness productivity, security and satisfaction.

The ISO/IEC 9126 quality model is a two level model as it consists of a set of characteristics (or attributes) and sub-characteristics. For instance, the internal and external quality characteristics are: Functionality, Reliability, Usability, Effiency, Maintenability and Portability; the sub-characteristics of Functionality are Suitability, Accurancy, Interoperability, Security and Compliance.

The ISO/IEC 9126 software quality model has been applied as follows. Initially a number of characteristics and associated sub-characteristics of interest for the project was identified. Subsequently requirements were grouped in 4 classes and then evaluated by determining whether each requirement implementation has or has not each sub-characteristic. Finally, a rating, in the range 0 to 10, has been computed for each requirement group and sub-characteristic by evaluating the ratio (expressed as a fraction of 10) between the number of the requirements in the group satisfying the sub-characteristic and the total number of requirements. The results of this evaluation are reported in Table 1.

Requirements	Adequacy	Accuracy	Compatibility	Learnability	Operability	Attractivity	Effectiveness	Productivity	Fulfillment
Formulation search									
Natural language search	■	■		■	■	■	■		■
Key concepts search	■	■	■		■	■	■	■	
Possibility of using logical operators	■	■						■	
Formulate queries in Italian	■	■	■	■	■	■	■	■	■
Formulate queries in English	■	■	■	■	■	■	■	■	■
Narrowing searches based on platforms, instrument, environmental matrix, geographic area	■	■	■	■	■	■	■	■	■
Rating	**10**	**10**	**7**	**7**	**8**	**8**	**8**	**8**	**7**
Presentation of search results									
Sort the results by relevance, date of publication, matrix environment, geography	■	■		■	■	■	■	■	■
Of each document display: title, author, source and internet link	■	■	■	■	■	■	■	■	
Document visualization	■	■	■	■	■		■	■	■
Rating	**10**	**10**	**7**	**10**	**10**	**7**	**10**	**10**	**7**
Refining search results									
Refining search on the base of file type (doc, pdf, etc)	■	■	■	■	■		■	■	
Refining search on the base of the author	■	■	■	■	■	■	■	■	■
Repeating one previous search			■	■		■	■	■	■
Rating	**7**	**7**	**10**	**10**	**7**	**7**	**10**	**10**	**7**
Averaged rating	**8.9**	**8.9**	**7.8**	**8.9**	**8.3**	**7.2**	**9.4**	**9.4**	**6.7**
Mean total rating	**8.4**								

Table 1. Identification of sub-characteristics in MAPS documents (see text above for details)

The total rating achieved is 8.4 out of 10, which is a remarkably good result. However, several areas have to be improved, especially is the presentation and usability extents.

In the second phase of the trial, the search engine was experimented to assess how much it was capable of retrieving relevant documents against the submitted queries.

Information Retrieval has elaborated several techniques to evaluate the performance of retrieval systems. Most of the methods use a defined collection of documents and a collection of queries for those documents; by comparing the query results provided by a group of experts with those returned by the search engine, it is possible to evaluate the performance of the engine. There are several techniques for comparing experts and search engine results, such as Precision and Recall, Accuracy, f-measure, ranked results, and so on. The major difficulty with these methods is linked with the need of experts and the fact that this may result in a subjective evaluation.

To overcome this difficulty, benchmarking techniques has been proposed, such as TREC (Text Retrieval Conference - http://trec.nist.gov/) and others. Basically, benchmarking techniques compare search engines using a well-defined collection of documents and queries. These techniques are not subjective as the involvement of experts is not necessary; however the collection of documents and queries are fixed and cross-domain in order to compare effectively different search engines.

Another set of techniques for evaluating the effectiveness of information retrial systems is based on evaluating ranked retrieval results, where importance is placed not only on obtaining the maximum number of relevant documents but also on returning relevant documents higher in the ranked list. A commonly used

method to rank outputs is to compute precision at various level of recall (see below for a definition of precision and recall).

Other measures have also been conceived to evaluate different information retrieval problems. For example, to measure the success of search tasks where just one relevant document is required, measures such as mean reciprocal rank (MRR – e.g. Chapelle et al., 2009), can be used.

Several techniques are then available for measuring the effectiveness of search engines (or information retrieval systems) and, in practice it is important to select an evaluation measure that is suitable for the given task.

We decided to adopt the Precision/Recall method for its simplicity and because benchmarking techniques, which involve the use of fixed document collections of several different domains, were not feasible as the MAPS search engine was specifically customized for the marine domain.

Precision and recall are defined as follows:

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

or using , using a mathematical notation

$$p = \frac{n_R^R}{n^R} \qquad r = \frac{n_R^R}{n_R}$$

where n^R is the number of document retrieved by the search, n_R is the number of relevant documents in the documents collection and n_R^R is the number of relevant documents retrieved by the search.

For computing precision and recall, a collection of 15 documents and an assortment of 9 queries (Q1, …, Q9 in Table 2) were defined; for each document-query pair, a number of experts determined the corresponding relevance (n_R in Table 2). Then the 15 documents have been loaded in an empty installation of the MAPS pilot system and the 9 queries have been executed, annotating the number of retrieved documents (n^R in Table 2) and the number of relevant documents (n_R^R in Table 2) retrieved by each query. Finally, overall precision and recall values have been computed by averaging the values resulting from each query. The results obtained are reported in Table 2.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9
# of Relevant docs (n_R)	2	2	4	2	2	2	2	4	1
# of Retrieved & Relevant docs (n_R^R)	2	2	2	0	0	2	2	3	1
# of Retrieved docs (n^R)	2	4	2	0	0	2	2	3	1
Precision ($p = \frac{n_R^R}{n^R}$)	1.00	1.00	0.50	0.00	0.00	1.00	1.00	0.75	1.00
Recall ($r = \frac{n_R^R}{n_R}$)	1.00	1.00	0.50	0.00	0.00	1.00	1.00	0.75	1.00
Precision (Average)	0.75								
Recall (Average)	0.81								

Table 2. Precision and Recall values evaluated in the trial (see text above for details)

By averaging the precision and the recall computed for each query, we obtained that about 80% of relevant documents were retrieved while 75% of the retrieved documents were relevant. Which is a promising result.

Further analysis shows that the behavior of the system is largely dependent on the marine domain customization (that we implemented by using a domain ontology). For instance, by restricting the ontology to the terms of the P02 BODC vocabulary, the system performs better with documents describing marine data acquisition procedures and protocols. Including additional concepts to the ontology, the system starts to perform also with other types of marine documents, such as articles and studies.

FUTURE RESEARCH DIRECTIONS

It is well-know that document indexing techniques are not sufficient to satisfy user information needs that go beyond the limits of a simple term matching search. Therefore the MAPS search engine is enriched with semantic technologies aimed at providing an accurate representation of the content of vast repositories of unstructured documents for semantic indexing purposes. Today language technologies make it possible for scientists and developers to produce software applications capable of revealing the semantic properties of textual elements and associating them with conceptual structures. Search functions coupled with semantic-conceptual technologies are able to interpret the underlying search criteria and thus better identify the data and the corresponding documents. This enables an effective and selective access to information even in the presence of significant collections of data.

CONCLUSIONS

The MAPS project, supported by regional funding POR-FESR for industrial development of enterprises associated to the Liguria Cluster of Marine Technologies, is part of a long term activity aiming at building a computer platform for supporting a Marine Information and Knowledge System. This integrated MAPS platform offers the advantage to have all content in one place and allows linking different information that exists, thus helping data management activities.

This study has attempted to provide semantic annotations to grey literature documents of the oceanography domain. The initial experiment has revealed that available methods are capable of assisting the process of semantic annotation with promising results. The incorporation of ontologies and knowledge resources in a rule-based information extraction technique promises to enable rich semantic indexing of grey literature documents.

Flexibility is a necessary requirement for any new use that the system intends to embrace while additional efforts are required for further exploitation of the technique and adoption of formal evaluation methods for assessing the performance of the method with measurable criteria.

REFERENCES

Bartolini, R., Lenci, A., Montemagni, S., Pirrelli, V. &Claudia Soria C. (2004) *Semantic mark-up of Italian legal texts through NLP-based techniques*, in Proceedings of LREC 2004, Lisbona: 795-798.

Belov, S. & Mikhailov, N. (2012). *Technical Workshop on the IODE Ocean Data Portal,*. Retrieved Oostende, Belgium, 27-29 February2012 Status of development of the ODP V2 from ftp://ftpdownload:ftpdownload@odp.meteo.ru/5-ODP_V2.pdf

Boehm, B., Brown, J.R., Kaspar, H., Lipow, M., G. McLeod, G.M. & M. Merritt, M. (1978) *Characteristics of Software Quality*, North Holland, New York.

Boyer, T.P., Antonov, J.I., Baranova, O.K., Coleman, C., Garcia, H.E., Grodsky, A., Johnson, D.R., Locarnini, R.A., Mishonov, A.V., O'Brien, T.D., Paver, C.R., Reagan, J.R., Seidov, D., Smolyar, I.V. & Zweng, M.M. (2013). *World Ocean Database 2013*. Sydney Levitus, Ed.; Alexey Mishonov, Technical Ed.; NOAA Atlas NESDIS 72, 209 pp.

Bosma, W. E., Vossen, P., Soroa, A., Rigau, G., Tesconi, M., Marchetti, A., Monachini, M. & Aliprandi, C. (2009). *Kaf: a generic semantic annotation format*. In Proceedings of the GL2009 Workshop on Semantic Annotation, Pisa, Italy.

BRIN, S. (1998). Extracting Patterns and Relations from the World Wide Web. In *Proceedings of the WebDB Workshop at 6th International Conference on Extending Database Technology*, EDBT'98. Pages 172-183.

Brown, J. S. & Duguid, P. (1998). *Organising knowledge*. California Management Review, 40,3, 90-111.

Buck, J. & Lowry,R. (2017). *Oceanographic Data Management; Quills and Free Text to the digital age and 'Big Data'*. In: Diviacco, P., Leadbetter, A., Glaves,H., Oceanographic and Marine Cross-Domain Data Management for Sustainable Development, Hershey, PA: IGI Global.

Chapelle, O., Metlzer, D., Zhang, Y. & Grinspan, P. (2009). *Expected Reciprocal Rank for Graded Relevance.* In *Proceedings of the Conference on Information and Knowledge Management*, CIKM'09, Hong Kong, China.

Creative Commons (2001) "Some Rights Reserved": Building a Layer of Reasonable Copyright. Retrieved 2005-09-14 from the World Wide Web: http://creativecommons.org.

Clarke, R. (2005). A proposal for an open content licence for research paper (Pr)ePrints. *First Monday 10* (8), 1-11. http://www.firstmonday.org/issues/issue10_8/clarke/.

Diviacco, P. & Leadbetter, A. (2017). *Balancing Formalization and Representation in Cross-Domain Data Management for Sustainable Development.* In: Diviacco, P., Leadbetter, A., Glaves, H., Oceanographic and Marine Cross-Domain Data Management for Sustainable Development, Hershey, PA: IGI Global.

European Commission (2010) *European Marine Observation and Data Network: Impact Assessment*. SEC(2010) 999 Final.

European Commission (2012) *Green Paper Marine Knowledge 2020*. COM(2012)473Final (http://ec.europa.eu/maritimeaffairs/documentation/publications/documents/marine-knowledge-2020-green-paper_en.pdf)

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Glaves, H. (2017). *Developing a Common Global Framework for Marine Data Management*. In: Diviacco, P., Leadbetter, A., Glaves, H., Oceanographic and Marine Cross-Domain Data Management for Sustainable Development, Hershey, PA: IGI Global.

Kogut, P. & Holmes, W. (2001). *AeroDAML: Applying Information Extraction to Generate DAML Annotations from Web Pages*. In: First International Conference on Knowledge Capture (K-CAP 2001). Workshop on Knowledge Markup and Semantic Annotation. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.21.8180&rep=rep1&type=pdf

Lenci, A., Montemagni, S. & Pirrelli, V. (2003) *CHUNK-IT. An Italian Shallow Parser for Robust Syntactic Annotation*, Linguistica Computazionale, 16-17: 353-386.

Lott, B. (2012). *Survey of Keyword Extraction Techniques*, UNM Education.

Maillard, C., Lowry R.K., Maudire, G. & Schaap, D. (2007) *SeaDataNet: Development of a Pan-European Infrastructure for Ocean and Marine Data Management*. Oceans2007, DOI**:**10.1109/OCEANSE.2007.4302435

Manzella, G. & Manzella, A. (2015). *Apeiron: engage students in earth and ocean sciences*. International Journal of Knowledge Society Research (IJKSR), IGI Global, 6, 4, 100- 111, DOI: 10.4018/IJKSR.2015100107.

Marinelli, R. & Roventini, A. (2006) *The Italian maritime lexicon and the ItalWordNet semantic database*. In: Linguistic in the twenty first century, Ed. E.M. Bermudez and L.R. Miyares, Cambridge Scholar Press, 173 – 181.

McCall, J.A., Richards, P.K. & Walters, G.F. (1977) *Factors in software quality.* National Technical Information Services (NTIS), RADC-TR-77-369.

Powers, D.M.W. (2011). *Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation.* Journal of Machine Learning Technologies **2** (1): 37–63.

Paul Clough, Mark Sanderson (2013), *Evaluating the performance of information retrieval systems using test collections,* Information Research, Vol. 18, No. 2 (http://www.informationr.net/ir/18-2/paper582.html)

Proctor, R., Roberts, K. & Ward, B.J. (2010). *A data delivery system for IMOS, the Australian Integrated Marine Observing System*, Advances in Geosciences, vol. 28, pp. 11-16, doi:10.5194/adgeo-28-11-2010.

Polanyi, M. (1966). *The Tacit Dimension*. University of Chicago Press: Chicago

Rajman, M., & Besançon, R. (1998). Text mining: natural language techniques and text mining applications, in Data Mining and Reverse Engineering , Springer US, 50-64.

Roventini A., Alonge A., Calzolari N., Magnini B., Bertagna F. (2000), "ItalWordNet: a Large Semantic Database for Italian", in *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, Athens, Greece, 31 May – 2 June 2000, Volume II, Paris, The European Language Resources Association (ELRA), 783-790.

UNESCO (2013). *Berlin Declaration*. http://www.unesco.org/new/en/social-and-human-sciences/themes/physical-education-and-sport/mineps-2013/declaration/

Vossen, P., Agirre, E., Calzolari, N., Fellbaum, C., Hsieh, S., Huang, C., Isahara, H., Kanzaki, K., Marchetti, A., Monachini, M., Neri, F., Raffaelli, R., Rigau, G., Tescon, M. & van Gent, J. (2008). *KYOTO: A System for Mining, Structuring, and Distributing Knowledge Across Languages and Cultures*, in: Proceedings of the Fourth International Global Word Net Conference - GWC 2008, Szeged, Hungary, January 22-25, 2008.

Wang, Y. (2008). *Software Engineering Foundations, a software science perspective*, Auerbach Publications.

Wannever, B. (1945) *As we may think*, The Atlantic Monthly (http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/)

Ward, J. & Peppard, J. (2002) *Strategic Planning for Information Systems* (3rd Edition), Chichester: Wiley

WCED (1987). *Report of the World Commission on Environment and Development: Our Common Future.* http://www.un-documents.net/our-common-future.pdf

KEY TERMS AND DEFINITIONS

Data: re-interpretable representation of information in a formalized manner suitable for communication, interpretation, or processing [ISO/IEC 2382-1:1993]

Information: knowledge concerning objects, such as facts, events, things, processes, or ideas, including concepts, that within a certain context has a particular meaning [ISO/IEC 2382- 1:1993]

Information system: any organized system for the collection, organization, storage and communication of information.

Information retrieval: the activity of obtaining information resources relevant to an information need from a collection of documents.

Lemmatisation: the process of grouping together the different inflected forms of a word so they can be analysed as a single item.

Part of speech: a category of words which have similar grammatical properties. Commonly listed English parts of speech are *noun*, *verb*, *adjective*, *adverb*, *pronoun*, *preposition*, *conjunction*, *interjection.*

Search engine indexing: collects, parses, and stores data to facilitate fast and accurate information retrieval

Tokenization: the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens. The list of tokens becomes input for further processing such as parsing or text mining.